DOCUMENT RESUME

ED 179 585 TM 009 875

AUTHOR	Burry, James, Ed.; Quellmalz, Edys S., Ed.
TITLE	Issues in Measurement and Methodology: CSE's 1978 Conference.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
PUB DATE	Nov 78
NOTE	17p.; Abstracts of papers presented at the Measurement and Methodology Conference of the Center for the Study of Evaluation (1978)
JOURNAL CIT	Evaluation Comment: v5 n4 Nov 1978
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Standards; Achievement Tests; Cognitive Measurement; Curriculum Evaluation; Data Analysis; Educational Policy; *Evaluation; Evaluation Criteria; *Evaluation Methods; Mastery Tests; *Measurement; Minimum Competency Testing; Models; Program Evaluation; Public Policy; Research Methodology; Research Problems; *Statistical Analysis; *Test Construction; Testing Problems; Test Selection
IDENTIFIERS	Elementary Secondary Education Act Title I; Project Head Start; RMC Models

ABSTRACT

Abstracts are presented of the major conference papers and thematic discussions delivered at the 1978 Measurement and Methodology Conference. The titles of the presentations are: Policy-Responsive Evaluation (Wiley); When Educators Set Standards (Glass); Comments on Wiley and Glass (Schutz); Key Standard-Setting Considerations for Minimal Competency Testing Programs (Popham); Evaluation of Title I via RMC Models--A Critical Review (Linn); Comments on Linn and Popham (Millman); Prediction Analysis and the Reliability of a Mastery Test (Wilcox); An Achievement-Test Item Model (Harris); A Head Start Re-evaluation: Positive Effects Are Not Yet Demonstrable (Bentler and Woodward); Analyzing Multilevel Educational Data: The Choice of an Analytical Model Rather than the Unit of Analysis (Burstein); Is Something Better than Nothing? Metaphysical Test Design (Baker); The Measurement of Expertise: Analysis of the Development of Knowledge and Skill as a Basis for Assessing Achievement (Chi and Glaser); Test Design: A View from Practice (Shulman); The Design of Measurement in Instruction (Schutz); The Demise of Generality in Measurement and Research Methodology (Frase); Don't They All Measure the Same Thing? Consequences of Standardized Test Selection (Floden et al); and Comments on Floden and Frase (Quellmalz). (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for the Study of Evaluation

November 1978 Vol 5, No 4

CSE EVALUATION COMMENT UCLA

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED179585

Issues in Measurement and Methodology: CSE's 1978 Conference

Edited by:
James Burry
Edys S. Quellmalz

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mark Young

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

1 **In this issue:**

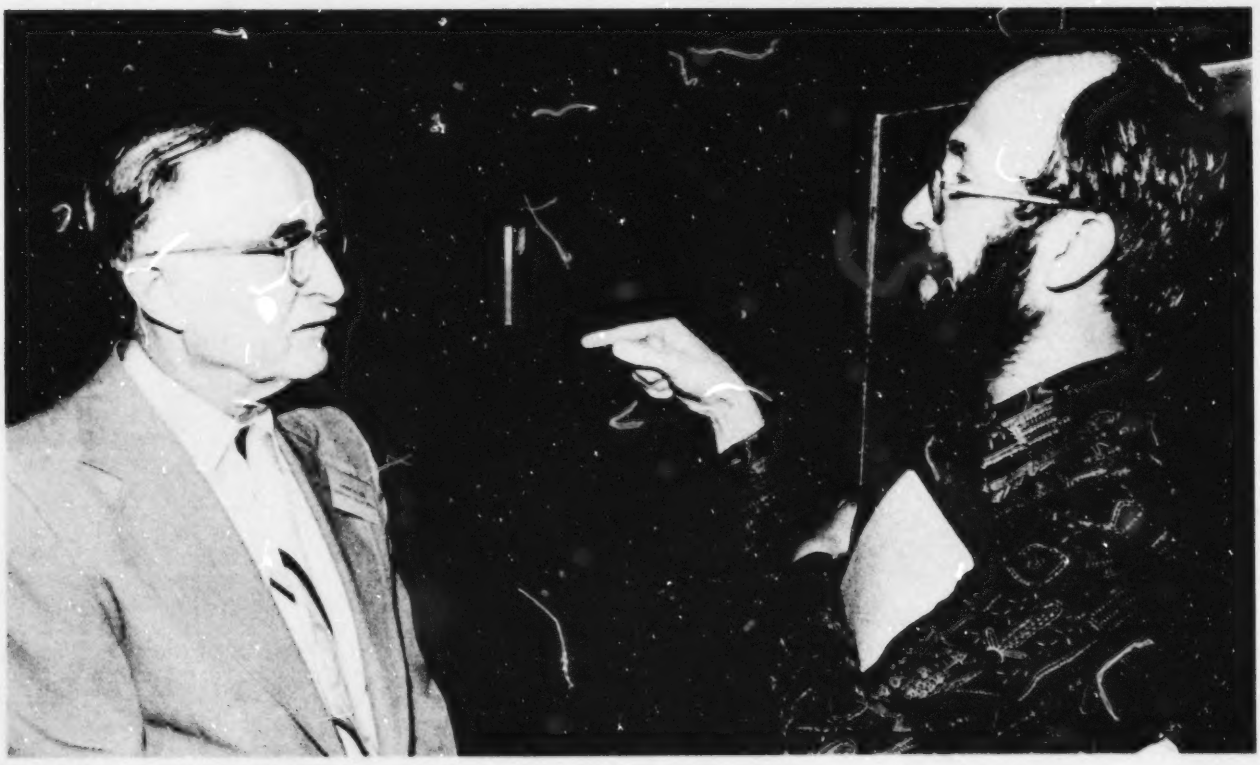
A series of presentations from CSE's 1978 *Measurement and Methodology Conference*. For this conference, CSE invited eminent scholars to address themes of *Measurement and Public Policy, Quantitative Theory and Application*, and *Test Design*. The major conference presentations as well as thematic discussions have been abstracted for this issue of *Comment*.

Addressing measurement and policy implications are: David E. Wiley on policy-responsive evaluation; Gene V Glass describing one state's difficulties in setting meaningful performance standards for minimal competence; W. James Popham on how such standards might be set; Robert Linn with a review of the RMC models used in Title I evaluations; and discussions by Richard E. Schutz and Jason Millman.

On the issue of quantitative theory and application are: Rand R. Wilcox on prediction analysis and reliability in mastery testing; Chester W. Harris discussing achievement test item models; P. M. Bentler and J. Arthur Woodward on Head Start and the question of positive effects; and Leigh Burstein addressing the implications of model selection in the analysis of multilevel data. A presentation by Frederick Lord on item-response theory for mastery testing is being published elsewhere and has not been abstracted for this issue.

Discussing issues in test design are: Eva L. Baker on key elements for domain-referenced testing; Michelene Chi and Robert Glaser on the measurement of expertise; Lee S. Shulman viewing test design from practice; Richard E. Schutz discussing the design of measurement in instruction; Lawrence T. Frase on the demise of generality in research and measurement methodology; Robert E. Floden, Andrew C. Porter, William H. Schmidt, and Donald J. Freeman on the consequences of standardized test selection; and discussion by Edys S. Quellmalz.

While these abstracts can only provide sketches of the conference proceedings, we hope they stimulate thought and continued dialogue with CSE. We expect to publish the complete proceedings at a later date.



TM009 875

Center for the Study of Evaluation
UCLA

Eva L. Baker
Director

EVALUATION COMMENT is published by the Center for the Study of Evaluation, University of California, Los Angeles. CSE is the only organized federally-supported agency devoting its total scope of work to the exploration and refinement of strategies in evaluation. CSE was established by the U.S. Office of Education in 1966 and is currently supported by the National Institute of Education. For 12 years, CSE has furnished the educational community with products and research which meet the increasing requirement for the evaluation of educational and social action programs.

Each issue of **Evaluation Comment** will discuss topics in educational evaluation by presenting articles on evaluation theory, procedures, methodologies, or practices. For each issue of **Comment**, CSE will request contributions on a specific topic in evaluation from recognized experts in the field; unsolicited manuscripts will not be accepted.

Evaluation Comment has a current readership of over 10,000 readers. Each person on our mailing list receives a free copy of **Comment**. Where additional copies are needed readers are encouraged to reproduce **Comment** themselves. To be placed on our mailing list, please write to:

James Burry, Managing Editor
Evaluation Comment
145 Moore Hall
University of California, Los Angeles
Los Angeles, California 90024

Editorial Board

Eva L. Baker	Chairperson
Tomas Arciniega	Dean, School of Education San Diego State University
William Johnston	Superintendent, Los Angeles Unified School District
John B. Peper	Superintendent, Anchorage School District
Michael Scriven	Director, Evaluation Institute University of San Francisco
Robert E. Stake	Director, CIRCE, and Professor of Educational Psychology at the University of Illinois
Daniel L. Stufflebeam	Professor of Educational Leadership and Director of the Evaluation Center, Western Michigan University

Measurement and Public Policy

Policy-Responsive Evaluation

David E. Wiley
University of Chicago and CEMREL, Inc.

Most of what passes for curricular or instructional innovation consists of modifying the goals and/or the resources for schooling and implementing these modifications in customary ways—i.e., by changing the content taught or the level of intensity with which instruction takes place. These two modifications directly affect the measurements resulting from achievement testing and when these effects are unaccounted for, they greatly distort interpretations of the data. These distortions explain much of the difficulty encountered in finding meaningful effects of innovations on achievement.

When an innovation is embedded in a school setting it replaces the program in use. Accordingly, new modal curricular emphases replace the old; this is one of the primary reasons for adopting innovations. However, there is also great variation in emphasis among implemented curricula, even those with the same program label. In the Follow Through planned variation, for example, the average amount of instructional time spent on reading and language ranges from 29 percent to 59 percent across selected curricula, but variation in emphasis within several of these is even larger.

Thus, most curicular innovations tend to involve significant changes in goals and, therefore, significant changes in *what* is learned, and these "qualitative" changes invalidate quantitative inferences concerning how much is learned. Consequently, I believe that conclusions concerning the relative effectiveness of alternative programs cannot be drawn using current methods.

A second achievement-affecting consequence of curricular innovations is a change in resource levels as a result of increased funds and regulations controlling their expenditure. Primarily, this effect is seen in lowered adult-to-pupil ratios, either by reduced class size, increased use of paid aides or parent volunteers, or the hiring of (e.g. reading) specialists. And these effects surely ought to influence the levels of pupil progression through the curriculum. We expect more learning for time spent with a child in a smaller group, receiving more individual attention. In addition, we would expect differential impact across subject areas. However, the effects of resource-level changes are very difficult to detect using achievement data as currently analyzed because of their association with curricular goal changes.

Detection of the effects of goal and resource level changes is also complicated by the great amount of natural variability among teachers in allotment of instructional time. In reading, especially, teachers have a lot of freedom in their organization of instruction and have different preferences. These differences in curricular emphases tend to be even larger than those planned. And when superimposed over the smaller curricular discrepancies, this within-group variation of teachers' preferences makes the achievement-relevant program and resource effects doubly difficult to detect.

Solutions to these problems are possible. But those solutions must explicitly take into account variations in curricular emphasis and must embody more sensible approaches to looking at the allocations of resources before we will ever be able to strip off gross effects for more subtle ones.

When Educators Set Standards

Gene V Glass
Laboratory of Educational Research
University of Colorado

Minimal competence cannot be determined in statistical or psychological ways; it can only be determined arbitrarily and with considerable risk of gross error and disruption. Because of this, minimal competence graduation testing will lead to situations in which many seniors may be failed unjustly; educators are, then, likely to be accused of not knowing what students need to know or else not teaching them what they need to learn. But educators are guilty on neither count; no one knows how well someone must read to succeed in life or what percent of a graduating class must be able to calculate compound interest payments.

This kind of situation is seen in Florida's law requiring high school students to pass an examination in basic skills and functional literacy. Students failing three times will be issued a certificate of attendance instead of a graduation diploma.

The Florida Department of Education was directed by the legislature to develop the test and determine a passing standard. Using objectives from previous assessment projects, the department contracted with ETS-Princeton to write the test items. Standards were set and the tests were administered; about 35% of the students failed the math exam and about 10% failed the reading and writing exam. It has been rumored that as many as 10% of the pupils in gifted programs failed the math test! The Florida Legislature appropriated $10 million for remedial programs; most of the funds will be allocated to math since the test seemed to show that is where most remedial work is needed.

The technical foundations of the program are surprisingly weak. When I interviewed a Florida Department of Education test specialist, I found that he was not consulted about the passing cut-off score; that criterion was set solely by the state Committee on Minimal Performance Standards. An interview with the chairman of the standards committee revealed that examination of the ETS items themselves was not the basis for setting standards, though it is doubtful whether that would have mattered; "mastery" was defined as 70% of the items correct since most of the Florida schools saw that as a "passing score." The chairman was not concerned that test items can be written harder or easier and still look about the same. Although the chairman recognized the arbitrariness of the difference in difficulty between the math and reading tests, he still advised schools on the basis of the scores to work on remedial math where the "incompetence" rates were highest.

An interview with the ETS director of the contract indicated that ETS has not been overly concerned with the possibility of different item writers developing easier or harder items for the same objective. The ETS manual on setting standards reflects a similar philosophy. Its recommended method of setting standards is egregiously untrustworthy; it also recommends that to avoid having the judgment of minimal competence based on test scores, judgments must be collected before score reports are distributed. With that, the manual precludes any hope for setting standards that might be safe and fair.

Florida's remedial education funds are being allocated on the basis of fatuous data generated by an indefensible assessment system. The system is supported by mistaken beliefs in a certain approach to measurement and standdards of minimal competence. A delegation of Florida educators and PTA officials has recommended that the state delay for one year the implementation of promotion and graduation on the basis of the functional literacy tests. A more appropriate recommendation would be immediate suspension of the minimal competence graduation requirements because the program is indefensible.

Comments on Wiley and Glass

Richard E. Schutz
SWRL Educational Research and Development

There is still a great gap between the measurement community and the public policy community. We in measurement and methodology are still engaging in a monologue that reflects *our* concerns rather than addressing the concerns of the public and its elected representatives. We are still not ready to begin a dialogue with the public policy community, though the motivation to begin that dialogue is certainly there.

Education today is both politicalized and technicalized; it will stay that way. While the integration of policy and science/technology in education is still in an early stage, the only question seems to be how it will occur, and not whether it will occur.

Wiley's intuitions are accurate. To evaluate what is learned we will have to look more closely at what is taught than we have done in the past. To find program effects, we will have to use indicators that reflect program implementation more closely than those we have traditionally relied on.

There is more verification behind Wiley's conjectures than those Wiley relies on. These data bases show that when there are sensible indicators of effects to track an educational program which has identifiable operational integrity, consistent functional regularities can be isolated with high reliability.

Glass's wish to stop minimal competency testing is out of step with the times. Minimal competency testing is here and the question is what to do about it. To make any positive contribution to education policy, we will have to learn how to help the school community deal with its realities, and not remain professionally content with criticism.

Contrary to the beliefs of many people in the research and development community and the school community, the public holds both communities in high regard. But this confidence can easily be eroded. If we continue to say that everything we stand for is bad, we can expect people to believe us, but we should not expect people to support our activities. The advance into the public policy arena brings new responsibilities to all of us in the measurement and evaluation community. Learning to deal effectively in this new arena will require augmentation of our present proficiency.

Key Standard-Setting Considerations for Minimal Competency Testing Programs

W. James Popham
University of California, Los Angeles
and Instructional Objectives Exchange

Educators are increasingly faced with the decision of determining how well high school students in the United States should perform in order to obtain a diploma. State legislatures and boards of education, more and more, are enacting programs that require students to pass minimal competency tests. One motive behind these minimal competency testing programs is to halt the devaluation of the high school diploma. If an excessively low standard is set, such devaluation will be accelerated rather than impeded and society is ultimately the loser. On the other hand, if an excessively high standard is set, then many students will be irrevocably penalized.

Some writers have argued that it is inappropriate to set performance standards for minimum competencies because these standards cannot be infallible. However, such testing programs, utilizing the most defensible standards possible, can and should be established. The standards they employ should be soundly based on relevant student performance and the expectations of various concerned groups. Three considerations should be kept in mind in the establishment of performance standards.

The first consideration is clarity of the target—the competencies for which the standards are to be set. By and large, norm-referenced (and some criterion-referenced) tests are too vague to provide the descriptive clarity necessary for making judgments regarding how well people ought to perform. Designers of minimal competency testing programs should work with a small number of highly significant kinds of competencies and rely upon criterion-referenced tests that clearly spell out what they are testing.

The second consideration is the use of test scores of examinees on the measures actually being used to assess student mastery of particular competencies. Standard-setters can be assisted in setting appropriate levels of performance if they have access to data from individuals who have been (1) uninstructed, (2) just instructed, and (3) previously instructed. By using data on how these kinds of individuals can peform a particular skill, standard-setters will be able to set more defensible standards for that skill.

The third consideration is the use of preference data from people who have a stake in the skill being considered and/or the style in which the performance is to be set; e.g., students, teachers, administrators, parents, business officials. These preferences will serve not to establish but to guide the performance standards set.

A great deal of work remains to be done related to establishing and trying out procedural variations for setting standards for minimal competency testing programs. Investigation should begin on such questions as how various "charges" given to standard-setters may influence the standards set; how differences in the clarity of a competency description may influence the kinds of standards set; how the differential prestige of the preference groups may influence standard-setters; how open versus closed standard-setting procedures may result in different kinds of standards; the degree to which standard-setting procedures can be systematized and/or quantified; the kinds of indicators that can be used to judge the reasonableness of standards set under varying procedural conditions.

Evaluation of Title I via RMC Models—A Critical Review

Robert Linn
University of Illinois at Urbana-Champaign

The three models which have been proposed, developed, and encouraged by RMC and the U. S. Office of Education for the evaluation of Title I programs have several shortcomings. The common goal of these models is to determine how much more students learn from participation in the Title I project than they would have learned without it. The principal concern of the RMC models is in obtaining the estimate of performance had the treatment not been implemented.

Model A, the norm-referenced model, assumes that posttest performance on a norm-referenced standardized test would be the same as pretest performance in the absence of treatment effects. Normal curve equivalents (NCE's) are used instead of percentiles for analyses and data aggregation. This model is low on rigor but high on feasibility. The actual comparison for Model A could be done in terms of either local or national norms.

A variation of this model is Model A2 which permits criterion-referenced tests or other non-normed measures to be used in place of normed, standardized tests. At pretest both the non-normed and norm-referenced tests are administered to provide an equating of the two, but at posttest only the non-normed test of interest is administered. The normed and non-normed tests must correlate at least .6. This is the least rigorous model.

Model B is the "control-group design," the idealized norm. It consists of random assignment of individuals to treatment conditions so that a control group is identified rendering the two groups comparable. This model may be highly rigorous, but it lacks feasibility.

Model C, the "special regression model," is basically the regression-discontinuity model where student assignment to Title I or a comparison group is based solely on the choice of pretest measure. This situation, as in Model B, is also relatively rare. Using analysis of covariance procedures, adjustments for pre-existing differences need to be made, but there are no guarantees that the adjustments will be correct. What is needed is the "complete covariate," a theoretically optimal composite of all conceivable variables on which the groups differ.

The adjustments used in Model A are based on Campbell's "fanspread hypothesis" that under similar conditions the groups fan out over time on commonly used scales such as grade-equivalents. In terms of NCE's the key assumption derived from the fanspread hypothesis is that, in the absence of a treatment effect, groups would maintain a constant NCE over time. Therefore, the test is to determine whether Title I intervention results in a shift in NCE scores over time.

Results are presented which suggest that the key assumption on which model A is based may not be valid. Certainly, the constant NCE assumption provides a weak foundation on which to build the principal model for the evaluation of Title I programs. Results are also presented which indicate that the aggregation of data from different tests may lead to unacceptably large variations. Finally, results are presented that show some serious limitations of Model A2.

All of the RMC models have serious drawbacks and fall short of yielding valid, comparable, and interpretable results. There seems to be a trade-off between rigor and feasibility. Although no solution is readily available, two

suggestions bear on the problem: (1) Educational evaluation is not mature enough to be ready for a fixation on a limited set of models; more flexibility and freedom to try different approaches is desired. (2) Project evaluations should not stand on the basis of one year's data. Historical data provide a better means of comparison than norm-referenced comparisons in light of the fact that norm groups may be inappropriate for the groups to be evaluated.

Comments on Linn and Popham

Jason Millman
Cornell University

Linn has pointed out that the RMC models neglect the importance of breadth of outcome measures, treatment definition, and description of the program's content. There are two additional points to raise. First, "no treatment" might be a bad label for the control group. Schools providing the norm also instruct, and many of them have Title I experience. The comparison is between the school's Title I program and some other program. Secondly, trying various approaches rather than fixing on a few models acts against the government's desire to aggregate data.

Popham's paper on standard setting makes a lot of sense. People who are against setting standards might use the following arguments. First, and perhaps the only reasonable one, is that standards are not needed if there are no decisions to be made. Second, and with little validity, is that minimums can become maximums. This concern can be reduced if competency testing is given early and frequently. Third, and totally unpersuasive, is that different standard setting methods might give inconsistent results. That is like saying that if one statistical test finds significance and the other does not, we give up on statistical tests. Fourth, and primarily attacking minimum competency standards, is we cannot say how much students need to know to succeed in life. But standards are not necessarily meant to imply what is needed for survival or the good life. Perhaps more neutral terms like "graduation requirements" could be used. Fifth, that standard setting is arbitrary. Popham handled this argument well by differentiating between "arbitrary" which implies randomness and "judgment" which can produce agreement among raters.

The question is not whether we need standards but how we establish them. Popham's plan for mixing relevant performance data with expert judgements seems reasonable. Finally, procedures for setting standards might be divided into different models depending on the inference the user wishes to make from the test performance.

Quantitative Theory and Application

Prediction Analysis and the Reliability of a Mastery Test

Rand R. Wilcox
Center for the Study of Evaluation
University of California, Los Angeles

A mastery test frequently consists of n dichotomously scored test items that represent some larger item domain. If x, the examinee's observed score, is greater than or equal to a specified passing score, x_0, he is advanced to the next level of instruction within a particular program of study; otherwise he is given remedial work. At a minimum it is desirable that such tests yield consistent decisions. In other words, if we pass (or fail) an examinee, we would like to have a reasonably high probability of making the same decision if we administer a randomly parallel, n-item test.

Recently there has been interest in estimating this probability or some function of it. Particular emphasis has been given to the proportion of agreement and also to Cohen's kappa. One approach to the problem is to assume that the conditional distribution of observed scores for a specific examinee is given by the binomial probability function. Let p be the probability of a correct response to a randomly chosen item. If we estimate p with x/n, as is typically done, we have an estimate of the binomial probability function which in turn gives us an estimate of the probability of the examinee getting y items correct on an n-item, randomly parallel test. We also have an estimate of the probability of $y \geq x_0$ and of $y < x_0$.

Recent results in the statistical literature suggest two possible alternatives for estimating a binomial probability function which in turn yield an estimate of the probability of passing (or failing) an examinee on a randomly parallel test. The first of these procedures stems from an attempt to approximate the Bayesian estimate of a probability function via a Taylor series expansion. The other approach is motivated by recent results that demonstrate, in terms of a commonly used goodness of fit criterion, the superiority of the Bayes solution for the case of multivariate normal distributions when a vague prior is used.

Using classical (i.e., non-Bayesian) statistical techniques, it is shown that this latter approach, when applied to the case of a binomial distribution, often gives substantially better results, particularly when n is small. Three criteria are used. The first is an ensemble squared error loss function involving the n + 1 probabilities associated with a binomial probability function. The second criterion is again based on squared error loss but this time we consider the estimation of $Pr(x \geq x_0)$. The last set of comparisons use a minimax approach.

A further consideration is the estimation of the proportion of agreement. In some cases, one of the new estimation procedures gives substantially better results than those obtained using more conventional techniques.

An Achievement-Test Item Model

Chester W. Harris
University of California, Santa Barbara

The paper describes a probability model for achievement test items which separates the conventional item difficulty (the proportion of students who correctly answer the item) into the product of two terms. The first term describes the probability of success for a typical examinee attempting the same task of which the item is a sample. The second term describes the probability of success (the probability of a correct response) in terms of the particular item that is being used. It is noted that this model can be used to estimate the proportion of the universe of such items that each student can handle. Moreover, the model recognizes that it is the combination of the difficulty of the task and the difficulty of a particular item that determines the overall difficulty or easiness of the item for the group, and thus the overall success of the group with this type of task.

A situation in which the model might apply is one where completion or production items are used although other types of items might fit the model as well. It is assumed that there is some undefined latent trait which determines the performance on an item for a typical examinee. Each student is assumed to belong to one of two categories associated with this latent trait, say + and -. For students in the + category, the second parameter of the model represents the conditional probability of failing the item. Finally, two situations are described which might yield an estimate of the two parameters of the model.

A Head Start Re-Evaluation: Positive Effects Are Not Yet Demonstrable

P.M Bentler and J. Arthur Woodward
University of California, Los Angeles

Reanalysis of Head Start data, using alternative causal models, provides no support for the hypothesis of positive impact of the Head Start program on the cognitive performance of its participants. The reanalysis includes a replication of work by Magidson, the application of an alternative procedure of testing for treatment effect, and the consideration of more complex models.

In a recent report, Magidson used a causal modeling approach towards evaluating the effect of a summer Head Start program on white first-graders' cognitive ability. Using what he called a new, general causal model which would permit an appropriate adjustment to be made for pre-existing differences among the participant groups, he reported that positive estimates of program effect could be shown for a sample which had yielded particularly negative effects in earlier analyses. This proposed "general alternative" to analyses of covariance (which turns out to be no more than a recommendation that Joreskog's LISREL model be used in ANCOVA situations) corrects any estimates of treatment effect for measurement errors in the context of a model that hypothesizes latent factors that influence overt variables and each other.

Posttest scores on two tests, the Illinois Test of Psycholinguistic Abilities and the Metropolitan Readiness Test, are used as dependent measures in the analyses. Independent variables include a treatment variable indicating participation in the Head Start program and four indicators of

socioeconomic status as control variables (mother's and father's education, father's occupation and income).

Our replication of Magidson's analysis shows an underlying structure of a general SES factor, two specific SES factors, and a fourth factor synonomous with the treatment; these and other results are similar for both dependent measures. Parameter estimates show a positive effect of the treatment factor on the true score estimates of the dependent measures. A chi square test shows an adequate model fit. These results led Magidson to the conclusion that the treatment had a small positive effect on both cognitive measures. However, standard errors for the treatment effect as large as the effect itself raise questions about such a conclusion.

Magidson's conclusions can also be questioned because of the procedure he used for analyzing the treatment effect. A better way of evaluating the treatment effect involves estimating the parameters of a second model that differs from the previous model only in that the treatment effect is fixed at zero and then comparing the two models via statistical methods. The difference in the chi square fit of the two models provides a significance test for the treatment effect. We apply this procedure and find that one cannot reject the hypothesis that the treatment effect on either dependent measure is zero (i.e., no evidence of a Head Start effect).

We also considered other more comprehensive models that include both dependent measures in a single analysis. Again, the comparisons of pairs of models, one with a treatment parameter and one with the treatment effect set of zero, result in non-significant chi square differences; that is, one cannot reject the hypothesis that the program effect is zero.

We have used Magidson's model and have developed more complex, similar models ourselves; we have found no evidence for positive Head Start effects using such models. Nonetheless, it is possible that a similar test carried out in the context of a completely different type of model might yield evidence favorable to the Head Start program.

Analyzing Multilevel Educational Data: The Choice of an Analytical Model Rather Than the Unit of Analysis

Leigh Burstein
University of California
and Center for the Study of Evaluation

The debate over the choice of the unit of anlysis has evolved into a more general methodological issue—how to analyze data derived from the multiple levels comprising an educational system. If the concern is to learn something about a particular question, then choice of analytical model may be more important than choice of the unit of analysis.

Selection of an analytical model is largely constrained by the study being conducted and the processes and outcomes being investigated. Experimental paradigms focus on power and significance rather than on the description, estimation, and explanation needed to study the educative process within the sociopolitical context of the classroom and the school. To derive the most tenable analytical models for educational research and evaluation, emphasis should be placed on quasi/nonexperimental methodology coupled with more intensive examination of within-classroom processes.

Most researchers facing the problem of analyzing multilevel data have tried, unsuccessfully, to choose among alternative units of analysis. Neither theory nor statistical considerations have been helpful in making these choices. Further, conflicting program foci—e.g., the child is the unit of evaluation but the classroom is the unit of treatment—have exacerbated the question of unit selection.

Different types of analysis of multilevel data address different questions, and research on schooling typically asks questions at multiple levels. When the substantive questions and/or statistical considerations warrant some analysis of aggregated data, the types of between-group effects remain to be specified. Further, if the purpose is to determine factors affecting pupil performance, analyses of between-group means may hide important differences in the within-group distribution of pupil outcomes and educational inputs.

Educational outcomes are multi-faceted and are incompletely measured by simple group averages. Thus, using distributional characteristics in addition to the mean as criterion measures has considerable merit for the following reasons: schools and classes differ in the distribution of educational performance as well as in mean performance; the effects of schooling on within-group behavior suggest examination of distributional properties other than group means; recent studies have shown empirically that distributional properties such as standard deviations and indices of skewness can add to the descriptive power of the mean.

A theory has been elaborated for using within-group slopes of outcomes on inputs as criteria in educational effects studies. The expectation is that, given similar distributions of entering pupil characteristics, different teachers and instructional practices will result in differing distributions of educational outcomes. Teachers who are equally effective in terms of means might yield varying slopes (within-class regression of outcomes on inputs) because of different instructional practices. A practice emphasizing improved performance of lower-ability students might produce a flatter slope, while a practice encouraging individuals to learn at their own rate might produce a steeper slope.

Previous studies have compared analytical models for identifying educational effects in the presence of hypothetical heterogenous within-group regressions. For the conditions simulated it was found that heterogeneous within-class slopes made important differences in the identified effects. Within-group slopes are therefore conceptually appealing indices of educational effects, but three points must be kept in mind. First, the slopes must be sufficiently stable. Second, the slopes must be potentially distinct from other group indices such as pre- and posttest means and standard deviations. Third, realistic cases must be found in which slopes are related to school and class characteristics after controlling for other background measures and group indices.

Recent investigations conducted at CSE provided empirical evidence of the relation of within-group slopes to school characteristics. In an analysis of science achievement data for 14 year-olds, the within-school slopes of science achievement on verbal ability was positively related to the mean responses of pupils within the school, to indices of exposure to science instruction, and to the degree to which students reported instructional practices which emphasized exploration (discovery methods of instruction). These results occurred despite controls for the school means and standard deviations of verbal ability and science performance and the means for pupil background measures. In contrast, school science achievement means and standard deviations were not strongly related to any school characteristics under consideration.

Test Design

Is Something Better Than Nothing? Metaphysical Test Design

Eva L. Baker
Center for the Study of Evaluation
University of California, Los Angeles

An area of great importance in testing but of strangely little effort is the area of test design. When one reviews the proportion of effort in psychometrics devoted to test design issues and contrasts it with work in the analyses of tests, one is astounded at the disparity. While periodic exhortations exist in the literature regarding elements of good test design, the research-supporting procedures for actually developing such tests are rather lightly treated. Of critical import, of course, is not that research-supported procedures exist, but that these procedures are developed with the understanding of the appropriate use of tests. Clearly, tests must be developed which share describable elements with instruction, so that the relationship between student performance on achievement measures on the one hand, and teaching and learning, on the other, are linked strongly.

The present technology for developing tests—from their specification, to determining the fit of items within a given strategy, to analysis and interpretation—is weak when one applies this criterion of instructional relationship.

Our research program at CSE in test design seeks to strengthen the relationship between testing, teaching, and learning, by exploring the use of tasks which not only can be translated from instructional to achievement situations but which connect as well to the real strategies used to learn and demonstrate competence. At CSE we have identified the field of writing as one where the importance of these relationships can be profitably explored. Our research effort in this area will address the issues of criteria for good writing, generalizability of writing skills within and among modes of discourse, task constraints as a means of influencing the writing process, and the impact of instruction on the whole lot.

Our program in test design is guided by four major principles:

1. That the dimensions to be tested should be public. Students, teachers, and parents should be able to understand expectations for performance, and test specifications should be produced which foster such understandings.
2. That testing should be much more economical, both of the student and teacher time it takes away from other activities, as well as in its administration, data processing, and interpretation requirements. There is much too much testing conducted out of habit and without serious intention of applying the findings to the direct improvement of services to learners.
3. That the testing should be related to instruction, and not float on a cloud above that critical activity. Tests need to be prepared so that their results can be conveniently and regularly used to revise teaching practices.
4. That testing itself should be an important and sensible activity from the learner's perspective. The items used in a test should have real meaning to the respondent group and the test should be a part of important learning experiences.

In testing we will continue to walk the balance between curriculum control and lack of relevance. But if tests can begin to demonstrate, through research, some of the principles identified above, the convictions which many of us share about the value of testing might become reality.

The Measurement of Expertise: Analysis of the Development of Knowledge and Skill As a Basis for Assessing Achievement

Michelene Chi and Robert Glaser
University of Pittsburgh
Learning Research and Development Center

Effective measurement of achievement requires that test developers have information about the psychological characteristics of advanced competence and expertise in the knowledge and skills taught in schools. The performance of the beginner is different from the performance of the skilled and knowledgeable individual. Knowledge about the changes from novice to expert performance would make it possible to devise achievement measures more directly focused toward efficient educational strategies—strategies that would take the novice through a different succession of stages, yield fewer failures of learning, and bring about the eventual attainment of higher levels of achievement.

Research on the development of competence in knowledge and skill can provide a theoretical and empirical basis for the assessment of performance. Studies would identify expert-novice differences and their implications for the measurement of achievement and the conduct of instruction. These studies could be carried out in terms of the two major interacting components of the human information processing system—content structure and cognitive processes.

Content can be investigated in terms of the quantitative and qualitative aspects of knowledge structures. To be an expert in a subject domain means that one knows more: having more central concepts or conceptual nodes in memory, more relations defining each node, more relations interrelating the nodes, more robust relations for retrieving related nodes.

Process differences are differences in cognitive strategies that are not immediately attributable to content knowledge, although they may be a function of differences in the knowledge base. Observable strategy differences between expert and novice that can provide focus for study are solution strategies, such as forward-backward search, means-ends analysis; planning strategies, such as setting goals, parsing the problem statement, seeking an isomorphic problem where familiar routines already exist, assembling processes for applying the operators of the problem; and self-inquiry, such as checking problem solutions, seeking possible errors, hypothesizing possible goals of the problem.

The essential educational task is to provide individuals with knowledge structures and cognitive process capabilities that are more susceptible to long-term development than are learning outcomes based upon education for short-run levels of competence. The intent is to obtain knowledge about the nature of these long-term changes so that the design of instruction and assessment procedures need not rely primarily on knowledge of short-term learning and achievement. One important implication of the qualitative changes in the long-term development of competence is that different modes of learning are active at different stages in the acquisition of skilled performance. Consequently, the level of performance needs to be assessed and then matched with appropriate instructional conditions in order for further expertise to be developed.

Lack of knowledge of the phases of development of complex performance can create artificial limits on performance that serve as detriments to the learning of advanced competence. Lack of knowledge of what to measure as an indication of expertise results in the use of criteria of performance on tests that may not adequately measure advanced capabilities. The long-range goal of the research should be to obtain knowledge about skills that are complicated enough to require years to achieve expertise, and then to theoretically and experimentally determine how such performance can be best measured, and thereby more effectively achieved and maintained for a greater number of individuals.

Test Design: A View From Practice

Lee S. Shulman
Michigan State University
and Institute for Research on Teaching

Practitioners in medicine and teaching share a common involvement with tests; yet there are differences in how, why, and with what degree of commitment practitioners in the two domains employ tests. Studies conducted in medicine offer information useful in curriculum and test development and shed light on the reasons for some of these differences. For example, physicians dealing with cases across different domains of internal medicine exhibit nongeneralizability; performance in one domain provides no basis for predicting performance in another case domain. It is possible to misconstrue the nature of the expertise by making unwarranted assumptions about the domain itself rather than doing systematic studies of how the expertise is used.

For example, if the expertise were domain specific, then perhaps medical students need not know all of biochemistry, physiology, anatomy etc., before doing clinical work. Instruction could be vertically organized to permit the beginning of clinical work earlier.

There is a correspondence between this and similar issues in education. For example, what if the expertise

represented in, say, the content of third and fourth grade performance tests, which is assumed to be prerequisite to fifth grade performance, did not represent the way a second grader became a fifth grade expert?

In medicine there seems to be a commitment to tests. After clinicians have been provided with the results of a fairly small number of tests, the diagnostic accuracy and quality of their plans does not increase. But although their diagnostic accuracy levels off, clinicians' confidence increases as they order more tests. Yet, why is it that teachers do not use or trust tests very much? The kinds of tests used are inconsistent with, and in many cases irrelevant to, the realities of teaching. Physicians do not treat a laboratory test as self-standing. They use observation, interview, touching and feeling, as well as laboratory testing. They then develop an assessment and a plan by aggregating across these sources of information.

But the strategy in education test development is to devise tests as portrayers of student performance, even though the test is only a small behavior sample. Perhaps we should treat tests as only one part of an assessment procedure and work on ways of helping teachers better use the other observations which they so frequently employ in the classroom.

Further, teachers do not focus on goals and outcomes. Teachers do not evaluate in terms of general assessments of achieved outcomes. They attend to variations in student activities; they think about content covered; they make decisions about pacing of instruction. Since tests do not measure anything relevant to these concerns, we should not be surprised that teachers find them of little value. A recent study has shown that the best proxy for achievement is "academic learning time," a combination of content covered and degree of engagement or involvement. The best predictor of ultimate student performance, then, is a combination of the two indicators that teachers already use to monitor student progress.

Test designers inhabit the interstices between two kinds of expertise. One kind of expertise is the achievement of pupils who are being educated; the other expertise is that of teachers who guide development of the first expertise. If learning is an inferred process, then teaching is rooted in the ability to use various cues to make inferences about that process so that learning can be guided. The well designed test can be a good tool to inform the expertise of the teacher in making necessary inferences about the developing expertise of the learner.

The Design of Measurement in Instruction

Richard E. Schutz
SWRL Educational Research and Development

Test design is no longer a technical obstacle for people concerned with education or educational research and development; the "how to" of test construction can be acquired by consulting any number of introductory texts.

But not all aspects of design and planning related to measurement and methodology have been wrapped up. The traditional measurement and methodology view is that if tests are designed and constructed properly, the only issue remaining is their use. The wise or appropriate use of tests is a problem that needs constant attention from measurement and methodology people, since school people and others often use tests unwisely or inappropriately.

This view is conceptually outmoded. Aspirations to get school people to use tests wisely stand little chance, given efforts to get standardized tests out of schools. Fortunately, an alternative paradigm and technical means are now available that remediate the conceptual deficiency and that are in synchronization with contemporary education.

The concern here is with measurement *in* instruction rather than measurement *of* instruction. In education and educational R&D the worlds of objectives, instruction, and measurement have operated as three separate worlds, with no functional links among them.

From the standpoint of measurement and methodology, the rationale for keeping measurement external to instruction has been that each instructional setting is different and each includes a near-infinite number of different instructional resource options. But there is strong evidence that around 90% of classroom time in elementary school reading and mathematics is referenced to common benchmark intentions and to a small number of textbooks in each subject. This means that measurement *in* instruction is highly tractable under this condition. While cursory examination of the instructional objectives and texts supports the common belief that there are only differences, extensive analyses shows that the big differences in the instructional indicators are *within* age/grade levels rather than *between* age/grade levels.

Instrumentation then, can be developed that school personnel and text publishers in general will regard as consistent with "their" program and that can be used in instructional accomplishments at the class, school, and district level. The instrumentation directly integrates measurement and instruction.

SWRL refers to this family of instrumentation as Instructional Accomplishment Information (IAI) systems. The measurement incorporated in IAI directly references what is being accomplished in a class, school, or district, and is a sound mechanism for acknowledging these accomplishments.

SWRL has also developed instrumentation within a system that embodies all the features of IAI and goes on to include resources for obtaining as well as acknowledging accomplishments. This instrumentation, Comprehensive Programs for Instruction (CPI), thus makes "high quality instruction" an operational rather than rhetorical matter.

A CPI product system has potential for both the school community and the R&D community. For example, with the CPI in reading, the R&D community can provide the school community with the resources required to teach virtually all children to read. CPI also pins down the benefits of criterion-referenced tests, and advances the formative evaluation/summative evaluation distinction. Summative evaluation need no longer be treated as a one-shot matter following completion of development. The CPI system measures provide sound feedback mechanisms that carry through with the treatment in its "natural" use. In turn, this permits longitudinal studies of instruction using different units (class, school, district) of analysis.

The Demise of Generality in Measurement and Research Methodology

Lawrence T. Frase
National Institute of Education*

We hear a lot about test score declines and proposals to institute competence tests. Since we do not understand the nature and extent of test score declines, we need to be thoughtful about the consequences of our testing proposals.

While we have tools for assessing academic success, these measurements are not tied to an analytic conception of how people learn and how they adapt to the demands of school tasks. We can predict success or failure, but we do not have the means to remedy potential failures. We need test items that stem from what we know about the skills that allow students to function in classroom activities. To do this, test development needs the support of cognitive and developmental expertise so that our capacity to build test items is matched by ability to translate these items into instructional methods.

We have previously outlined a model of the skills entailed in reading study activities. The model included task elements, a memory system, and specific processing activities that operate on items in memory or in the task. We found that the performances of encoding, rehearsal, integration, retrieval, and relation are implicated in a variety of learning tasks and test-taking activities. If we can identify stimulus characteristics that support specific intellectual performances, we can use them to write test items that impose clearly defined processing requirements on subjects; these might also be linked to instruction.

The model can be thought of as representing two broad properties of intellectual performance—building an internal data base upon which to operate (the representational problem), and the operations performed on that data base. When a person selects a multiple-choice answer on a test, we often assume that the answer is related to the information on the page in support of the alternative. This may not be true, and if not, it limits our ability to draw inferences about performance from test items. To the extent that a test-taker imports simplifying strategies into the test-taking situation, we do not know what data base provides the context for performance.

Our work with syllogistic reasoning has looked at two sub-problems of representation: partial representation, which is concerned with whether all relevant features of a sentence are or need to be encoded; and optional representation, which is concerned with the form of representation, given that a proposition is completely represented. If we can clarify the contributions of partial and optional representation in testing, we will learn some useful facts about subjects' adaptive strategies in test taking. For example, if we find that partial representation is a major cause of error, we can concentrate on subjects' attentional processes. If we find that optional representation is a more potent factor, we might shift our attention to the semantic and interpretive dimensions of subjects' activities.

The data base presents another problem. When subjects have sufficient information to solve a problem and the skills to manipulate the information, they may still fail. On the other hand, when they have insufficient information, they may succeed. The notion of the text data base as a necessary or sufficient condition for performance must therefore be approached cautiously. In short, we must explore our conceptions of cognitive processing, and also the adaptive behaviors of test-takers that simplify and convert nominal stimulus materials into something other than they are intended to be.

*Now at Bell Laboratories.

Don't They All Measure the Same Thing? Consequences of Standardized Test Selection

Robert E. Floden, Andrew C. Porter,
William H. Schmidt, and Donald J. Freeman
Institute for Research on Teaching
Michigan State University

A great deal of attention has been called to the match between the content of instruction and the content of tests. If the test content does not match instructional content, test results may give a distorted picture of the effects of instruction.

While the effect of test content on test scores has been acknowledged, the effect of test content on instruction has had little attention. Although a testing program may have global effects on instruction, its subtler influences on instructional content need further investigation.

Most districts administer norm-referenced tests to assess district improvement. While little is known about the criteria used to choose a test series, factors probably include cost, ease of administration and scoring, and reporting format. The content covered by each test is a criterion given insufficient attention.

School personnel have been led to believe that major test series differ little in topics tested. But these tests do not all measure the same thing. Four test series dominate the market in elementary reading and mathematics: Stanford Test of Basic Skills (SAT); the Iowa Test of Basic Skills (ITBS); the Metropolitan Achievement Tests (MAT); and the CTB/McGraw-Hill Comprehensive Tests of Basic Skills (CTBS). We have compared the content coverage of these series in fourth grade mathematics. For some operations, the tests are quite similar in the percentages of items they include. But there are differences too: 21% of the items on the MAT involve addition, which is about 8 percentage points more than the other tests; the ITBS has at least 5 percentage points fewer multiplication items than the other tests; grouping is tested by the SAT but not by the MAT or the CTBS. With respect to their materials, 6% of the items on the CTBS involve percentage, while the ITBS and MAT have no such items. The MAT and SAT both have items on alternative number systems, while the ITBS and CTBS do not.

Even when tests have been empirically equated, the choice of test can greatly influence the estimate of program success. Differences in test content may result not from test publishers' failure to include the appropriate content, but from a general lack of agreement on the definition of basic skills.

In addition to influencing estimates of program success, it is generally believed that testing programs influence the content teachers present—that they present content that follows content of the test. Those who believe that tests represent only a fraction of the important content and those who believe teachers should determine their own instructional content find this phenomenon undesirable. Those who believe the content of instruc-

tion should be uniform see testing programs as valuable for determining instructional content. Both groups base part of their assessment of tests on the assumption that teachers "teach to the test;" this assumption warrants further investigation. There are other factors that might influence teacher choice of content; among these factors are textbooks, teacher conceptions of subject matter, teacher assessment of student achievement, and student interest in subject matter.

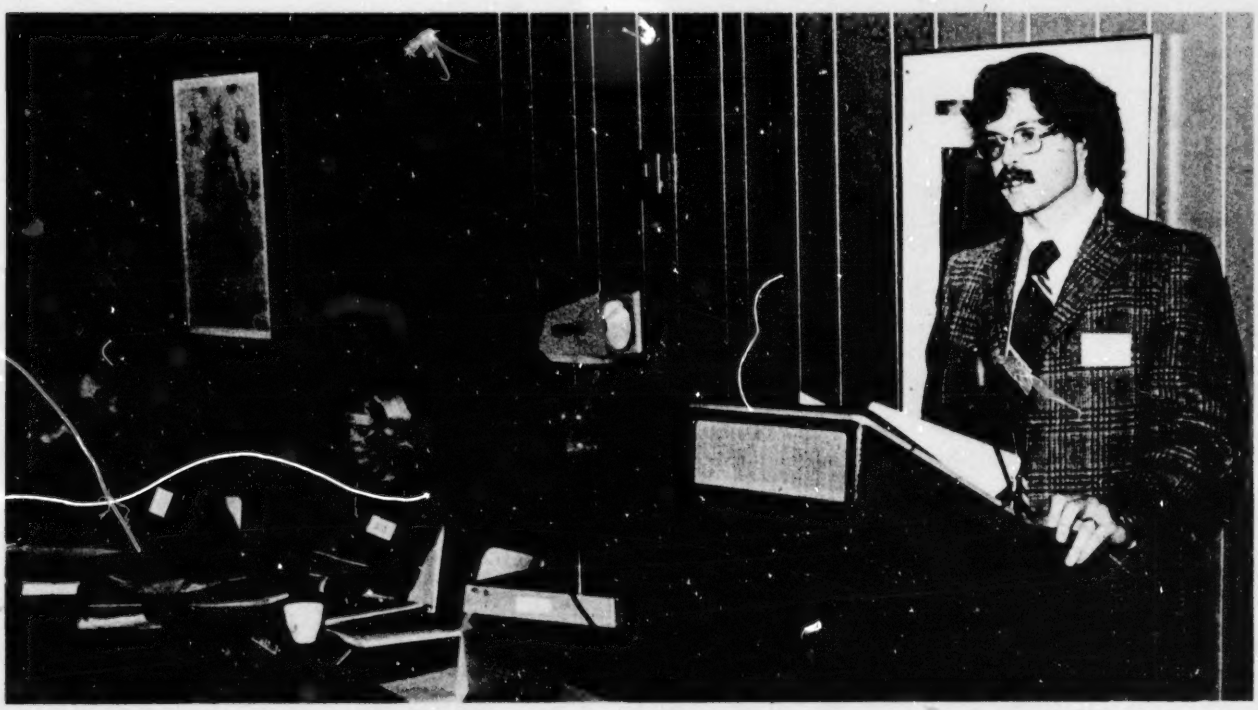
The existence of many possible influences on content choice implies a minimal requirement for district controls over instructional content in the classroom. Further, even if tests are found to have little influence on instructional content, it is still important that the content covered by the test match the content of most concern to those using the test results.

Comments on Floden and Frase

Edys S. Quellmalz
Center for the Study of Evaluation
University of California, Los Angeles

Floden and Frase continue themes previously raised: problems in state-of-the-art, norm-referenced test design and alternatives to current practice. Floden's paper reiterates the caution of other researchers that presumably equivalent norm-referenced tests for a particular subject area do not equally represent the same content. But the surface level of analysis and the content covered should be augmented by deeper analyses of the stimulus properties and response demands that contribute to differences in difficulty level in a domain within and between grade levels. Test publishers need to understand the importance of descriptive rigor and of models for communicating test design blueprints to a variety of concerned audiences.

In his concern with design of criterion-referenced tests for the more specific requirements of classroom diagnosis and prescription, Frase suggests that cognitive psychology can offer a framework for analyzing information processing demands of test tasks. It is important for those in test design to consider cognitive skills as a particular processing domain to assess, and to consider processing as a critical dimension in formulating subject matter achievement tests. Glaser's descriptions of stimulus and response dimensions differentiating novice and expert performance may provide a reference for constructing items on a continuum of difficulty. Frase's discussion of dimensions affecting students' retrieval of information and routines directly relates to test task features which can be incorporated into test specifications. It is likely that tests, rather than remaining curriculum free, globally structured, and superficially described, will become increasingly instructionally referenced, systematically formulated, and specifically characterized.



CSE COMMENTS

CSE Publications

CSE Reports—1978

Skager, R. W. The Great Criterion-Referenced Test Myth. No. 95.

Alkin, M. C. The Title of This Speech is . . . No. 96.

Alkin, M. C. A New Role for Evaluators. No. 97.

Wilcox, R. R. Some Comments on Selecting the Best of Several Binomial Populations or the Bivariate Normal Population Having the Largest Correlation Coefficient. No. 98.

Wilcox, R. R. Estimating True Score in the Compound Binomial Error Model. No. 99.

Wilcox, R. R. Estimating the Likelihood of False-Positive and False-Negative Decisions in Mastery Testing: An Empirical Bayes Approach. No. 100.

Wilcox, R. R. Applying Ranking and Selection Techniques to Determine the Length of a Mastery Test. No. 101.

Wilcox, R. R. Comparing Examinees to a Control. No. 102.

Walker, C. B. Standards for Evaluating Criterion-Referenced Tests. No. 103.

Burstein, L., Keesling, J. W., Conklin, J., & Doscher, M. L. Auditing a Large-Scale Evaluation: The Quality of Evaluative Information for the Assessment of Program Impact and for Decision Making. No. 104.

Burstein, L. Secondary Analysis: An Important Resource for Educational Research and Evaluation. No. 105.

Baker, E. L. Fun and Games: Their Relationship to Basic Skills in Elementary School Children. No. 106.

Dotseth, M., Hunter, R., & Walker, C. B. Survey of Test Selectors' Concerns and the Test Selection Process. No. 107.

Walker, C. B. Control Items: A Baseline Measure for Evaluating Achievement. No. 108.

Fischer, K. B. The Bilingual Prediction Project: Technical Report of the 1976 Pilot Study. No. 109.

Smith, L. S. Decisions and Dilemmas in Constructing Criterion-Referenced Tests: Some Questions and Issues. No. 110.

Conklin, J. Some Methodological Issues in Using Standardized Test Scores to Evaluate Large-Scale Educational Programs. No. 111.

Wilcox, R. R. Comparing Normal Populations to a Control. No. 112.

Porton, V. M. Grammatical Errors and Communication Breakdown. No. 113.

Millman, J. Determinants of Item Difficulty: A Preliminary Investigation. No. 114.

Wilcox, R. R. Estimating the Parameters of the Beta-Binominal Distribution. No. 115.

Malamuth, N. M., & Fitz-Gibbon, C. T. Tutoring and Social Psychology: A Theoretical Analysis. No. 116.

Fitz-Gibbon, C. T. An Analysis of the Literature on Cross-Age Tutoring. No. 117.

Fitz-Gibbon, C. T. A Survey of Tutoring Projects. No. 118.

Irizarry, R. Characteristics of Bilingual Programs Under Federal and State Legislative Policy: Implications for Program Development and Evaluation. No. 119.

Cabello, B., Elvenstar, D., Garcia, P., & Sisson, B. Variables Relevant to the Evaluation of Bilingual Education Programs: A Review of the Literature. No. 120.

Fitz-Gibbon, C. T. Tutoring: Some New Ideas. No. 121.

For information about ordering, contact Mark Young at the Center for the Study of Evaluation, UCLA Graduate School of Education, Los Angeles, CA 90024

CSE Evaluation Monographs—1978

#8 — Toward a Methodology of Naturalistic Inquiry in Educational Evaluation

In his monograph, E. Guba makes the case for naturalistic inquiry as distinguished from experimental inquiry. He then examines three major conceptual problems in doing evaluations using naturalistic methodology: boundary problems, focusing problems, and problems of authenticity. In so doing he summarizes work from other fields and demonstrates their applicability to evaluation.

#7 — The Logic of Evaluative Argument

The monograph by E. House takes the view that evaluation, lacking as it does the certainty of proof and conclusiveness, is fundamentally an act of persuasion. The analytical schema of the monograph grows out of a careful treatment of the premises of argument and agreement, distinguishing between the class of the "real"—facts and presumptions claiming validity vis-a-vis the universal audience, and the class of the "preferable" which includes values adhered to by a particular audience. Dr. House maintains that evaluation that is not mere oratory nor entirely arbitrary can greatly increase the possibilities for informed action.

#6 — Achievement Test Items—Methods of Study

This monograph examines methods of studying and characterizing certain types of achievement measures. The authors, C. Harris, A. Pearlman, and R. Wilcox, provide a model for describing student response data. They then review and critique the many measures of association and agreement which might be used for studying the equivalence and stability of achievement test items.

CSE Special Reports—1978

Setting up and Evaluating Tutoring Projects by Carol Taylor Fitz-Gibbon

Through a presentation of step-by-step planning questions, this report deals with the process of setting up and evaluating a tutoring project. The practical, systematic approach set forth in the report grows out of the author's experience as a teacher and out of her extensive research on cross-age tutoring as a compensatory strategy.

Characteristics of Bilingual Programs Under Federal and State Legislative Policy—Implications for Program Development and Evaluation by Ruddie Irizarry.

This report provides a review of federal and state bilingual education laws and an across-states summary which identifies the important variables to be considered in documenting and evaluating a bilingual program. Fourteen categories—such as population to be served, staff training requirements, student assessment, and evaluation reporting requirements—are used in organizing the summaries.

This report is available from the National Dissemination and Assessment Center—ESEA, Title VII, California State University, Los Angeles, School of Education, 5151 State University Drive, Los Angeles, CA 90032.

CSE COMMENTS

CSE Publications (Cont.)

CSE Test Evaluation Series

Educators continue to use the *CSE TEST EVALUATION SERIES*, a consumer's reference guide to standardized, published tests. Each volume contains a compendium of tests which have been evaluated for such characteristics as meaningfulness in relation to a specific goal, appropriateness for the intended student, usability for administration and decision making, and overall quality of standardization and technical excellence. The tests were judged by CSE staff trained to use a carefully devised evaluation procedure.

Each volume in the series focuses on a particular level of education for which a comprehensive list of educational goals has been developed by CSE. Superintendents, principals, testing officers, and other educators report that the CSE test evaluation books provide valuable assistance in selecting new test instruments and in the critical evaluation and revision of existing testing programs. The potential user need not have technical expertise in educational measurement and evaluation.

The following volumes are currently available:

CSE SECONDARY SCHOOL TEST EVALUATIONS
(1974—3 volumes, 900 pages, softbound 9" by 12")

CSE ELEMENTARY SCHOOL TEST EVALUATIONS
(1976—660 pages, softbound 9" by 12")

CSE-RBS TEST EVALUATIONS: TESTS FOR HIGHER-ORDER COGNITIVE, AFFECTIVE, AND INTERPERSONAL SKILLS
(1972—352 pages, softbound 9" by 12")

CSE Visiting Scholars Program

Nominations are being solicited for the 1979 CSE Visiting Scholars Program. This program, supported by the National Institute of Education, has been in existence for the past two years. Each year, we invite outstanding individuals from both the academic and practitioner communities to join us for varying amounts of time. Depending on schedules, CSE can support the work of visiting scholars for anywhere from a few weeks to several months. Our major activities, centering on the theory and practice of evaluation, are conducted in three program areas—evaluation and decision making, measurement and methodology, and bilingual education. Our field services and dissemination units assist in the utilization of research from the three programs. Visiting Scholars may either pursue a problem of their own choosing related to our scope of work or become involved in some aspect of CSE's activities. We welcome your nominations of individuals we should consider as Visiting Scholars. These nominations should be forwarded to Adrianne Bank, CSE's Associate Director.

Visitors to CSE

On September 7-8 an Advisory Panel convened at CSE to review the work done to date on the Bilingual Evaluation Technical Assistance project (under the direction of Ruddie Irizarry) and to provide advice which will help the project staff develop materials sensitive to the needs of bilingual educators. Members of the Panel included the following: Rudy Cordoba, Joel Gomez, Judith Guskin, Charles Leyba, Olivia Martinez, Antoinette Metcalf, Vivian Portillo, Egla Rangel, Susan Talley, Jose Vazquez, Anthony Vega, and Carl Wisler.

Recent CSE Activities in Addition to Ongoing, Federally Funded Programs

The California State Department of Education has awarded a five year contract to CSE to evaluate the School Improvement Program and Local Staff Development and Resource Centers Programs authorized by Assembly Bills 65 and 551. These laws call for improvement of California schools through comprehensive planning and implementation efforts and the support of these local efforts by school districts.

CSE is designing a formative evaluation system for the National Diffusion Network as part of a three year contract awarded to the Network (Andover, Massachusetts) to study federal and state dissemination strategies. After reviewing existing formative evaluation capabilities, CSE will develop training materials and a technical assistance strategy.

CSE is participating in the first year desegregation efforts of the Los Angeles Unified School District. The areas of CSE involvement include the following:

- CSE Staff and Faculty Affiliates are members of the Internal Evaluation Committee comprised of representatives from colleges and universities in the Los Angeles area.
- A CSE Faculty Affiliate is a member of the Court-appointed Monitoring Committee.
- A CSE Faculty Affiliate is a Court-appointed expert in the area of bilingual programs and their realtionship to desegregation.

Evaluation Field Services

CSE Workshops, Kits, and other resources from our research projects provide the bases for the training and technical assistance offered by Field Services. During the past year we have provided evaluation services to governmental agencies, school district, and schools throughout the United States. If you are interested in the range and costs of services available, contact:

James Burry, Director, Field Services
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, CA 90024
(213) 825-4711

CSE COMMENTS

CSE Welcomes New Staff

Donald W. Dorr-Bremme comes to UCLA from Harvard University where he taught in the Graduate School of Education, the institution from which he will shortly receive a Doctor of Education degree in Teaching, Curriculum, and Learning Environments. His previous work includes a variety of teaching and consulting in the public schools both in the East and in California. His special areas of interest are educational anthropology and socio-linguistics. Mr. Dorr-Bremme will be working on the CSE evaluation of the California School Improvement Program (AB-65). Specifically, he will oversee the intensive site studies to be conducted using ethnographic methodology.

Robert Wenkert will be dividing his time between teaching courses in field-work methodology in the UCLA Graduate School of Education and representing CSE, along with Adrianne Bank, on the Los Angeles Unified School District's Internal Evaluation Committee to oversee the first year's desegregation effort. Dr. Wenkert received a Doctor of Philosophy degree in Sociology from Oxford University, England. He comes to UCLA from the Bay Area where he served as a Senior Program Associate at the Far West Laboratory for Educational Research and Development, San Francisco, and taught at the University of California, Berkeley.

Antonio de Porcel, who will shortly be receiving his Ph.D. from Stanford University, joins the Center staff as the project director for the Bilingual Prediction Study. He has most recently been at CTB/McGraw-Hill in Monterey where he was an assistant project director responsible for reviewing new criterion-referenced tests in reading. Prior to that he worked at the American Institutes for Research and at the Stanford Research and Development Center. He has had extensive experience in evaluating bilingual programs and has a strong theoretical and practical background in design, analysis, and testing.

The first complete step-by-step guide to conducting a program evaluation...

Program Evaluation Kit

by
Lynn Lyons Morris
Carol Taylor Fitz-Gibbon
Marlene E. Henerson

Sponsored by the Center for the Study of Evaluation
University of California Los Angeles
Graduate School of Education

Most published work in evaluation is heavy on theory and short on procedures that the practitioner can actually use. The *Program Evaluation Kit* answers questions *procedurally*, as though it were a consultant called in to help you conduct your evaluation.

Each of the *Kit*'s eight books deals with one topic of practical concern. The *Evaluator's Handbook* is the nucleus of the *Kit*. Aimed primarily at practitioners who have never before conducted an evaluation of any size, the book contains step-by-step guides to help with planning and managing the evaluation. During this process, the evaluator reaches myriad decision points where one must assess the demands of the particular situation to choose among various courses of action. At each of these junctures the step-by-step guide recommends one of the *Kit*'s seven *How To* books. These give important information, in a common sense language, to consider in making the decision. Once the decision has been made, the books provide directions for following it through.

The *Kit* is an outstanding resource for evaluation, clearly written and easy to understand. It has been field tested at over 150 sites and used in university classes. Although the inexperienced evaluator will probably rely heavily on the Handbook, each of the *How To* books can be used separately.

The *Kit* was designed and written under the auspices of the Center for the Study of Evaluation at the UCLA Graduate School of Education with grant support from the National Institute of Education.

October 1978 8 volumes 1088 pages ISBN 0-8039-1073-8 $49.95
($55.90 if volumes ordered separately)

SAGE Publications, Inc.
P. O. Box 5024
Beverly Hills, California, 90210

Order Form

CSE Test Evaluation Series

Volume(s) of the CSE Elementary School Test Evaluations ..... @ $12.50 each
Set(s) of the CSE Secondary School Test Evaluations @ $22.00/set
Gr. 7&8 Gr. 9&10 Gr. 11&12 ..... or $8.00 volume.
Volume(s) of the CSE-RBS Test Evaluations: Tests for Higher-Order Cognitive, Affective, and Interpersonal Skills ..... @ $8.50 each

Total $ _____

CSE Monograph Series in Evaluation

Copy(s) of #1 — Domain-Referenced Curriculum Evaluation, W. Hively, et al., 1973 ..... @ $3.50 each
Copy(s) of #2 — National Priorities for Elementary Education, R. Hoepfner, et al., 1973 ..... @ $3.50 each
Copy(s) of #3 — Problems in Criterion-Referenced Measurement, Editors: C. Harris, M. Alkin, W. J. Popham, 1974 ..... @ $3.50 each
Copy(s) of #4 — Evaluation and Decision Making: The Title VII Experience, M. Alkin, et al., 1974 ..... @ $3.50 each
Copy(s) of #5 — Evaluation Study of the California State Preschool Program, R. Hoepfner & A. Fink, 1975 ..... @ $3.50 each
Copy(s) of #6 — Achievement Test Items: Methods of Study, C. Harris, A. Pearlman, & R. Wilcox, 1977 ..... @ $4.50 each
Copy(s) of #7 — The Logic of Evaluative Argument, E. House, 1977 ..... @ $4.50 each
Copy(s) of #8 — Toward a Methodology of Naturalistic Inquiry in Educational Evaluation, E. Guba, 1978 ..... @ $4.50 each

Total $ _____

CSE Special Report

Copy(s) of — — Setting Up and Evaluating Tutoring Projects, C. T. Fitz-Gibbon, 1978 ..... @ $3.00 each

Send order to: Office of Publications
CSE – UCLA
Graduate School of Education
Los Angeles, CA 90024

☐ My Payment is Enclosed for This Amount

☐ Please Bill Me for This Amount

Add $.50 for Postage and Handling $ .50

Calif. Residents Add Sales Tax $ _____

Grand Total $ _____

(Orders of less than $5.00 must be prepaid.)

CSE EVALUATION COMMENT

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024